

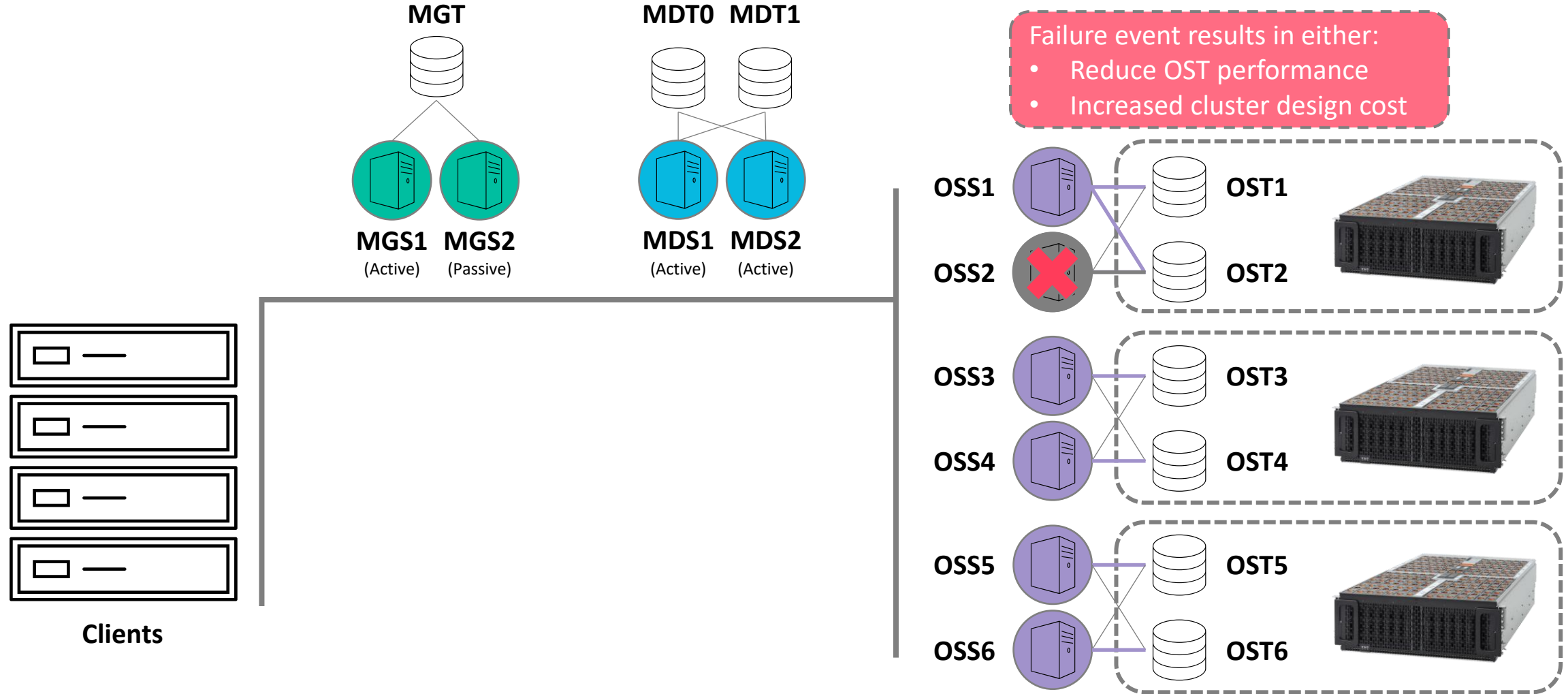


Improving Parallel File System Performance & Reliability with NVMe-oF™ Storage

Jonathan Flynn

Senior Technologist, Field Engineering, Platforms

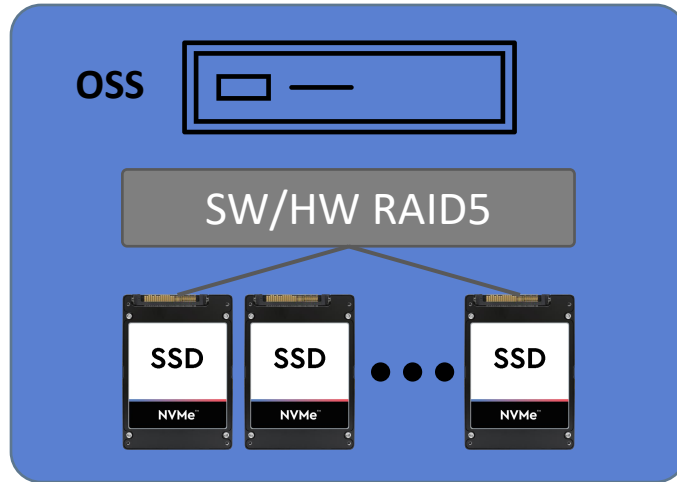
Traditional Parallel File System Architecture



Enclosure Evolution Challenges

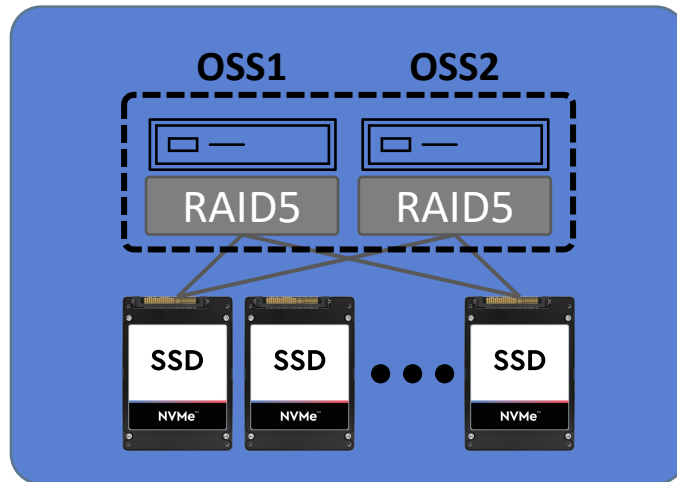


Current Approaches to NVMe™ Based OST



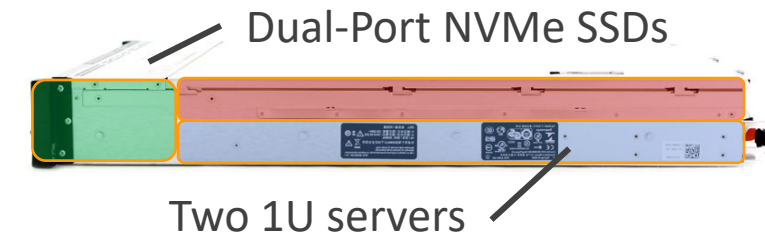
Single Node with Local NVMe

- Single OSS Node with local NVMe SSDs with RAID5/6 protection
- RAID protection provided by host SW (mdadm, erasure coding) or NVMe HBA
- No protection against OSS Server failure
- Limited performance using SW/HW RAID



HA Server with 24 Local Dual-Port NVMe

- Dual OSS Nodes w/ local Dual-port NVMe SSDs w/ RAID5/6 protection or erasure coding
- RAID protection provided by host SW (mdadm, erasure coding) or NVMe HBA
- Active/Active design limits max # of SSDs to 12 per OST or requires multiple NVMe name spaces.
- Traditional 8+1 or 8+2 base 2 OST RAID layout limits useful drives to 20 of 24



2U24 NVMe-oF JBOF

Enabling Standard NVMe PCIe® SSDs to be Shared in an External Enclosure

Enclosure

- 2U enclosure with dual IO Modules for HA
- Similar design to existing SAS SSD enclosures
- 24 standard dual-port NVMe PCIe SSDs
- No data services (i.e. no RAID). Just pass through

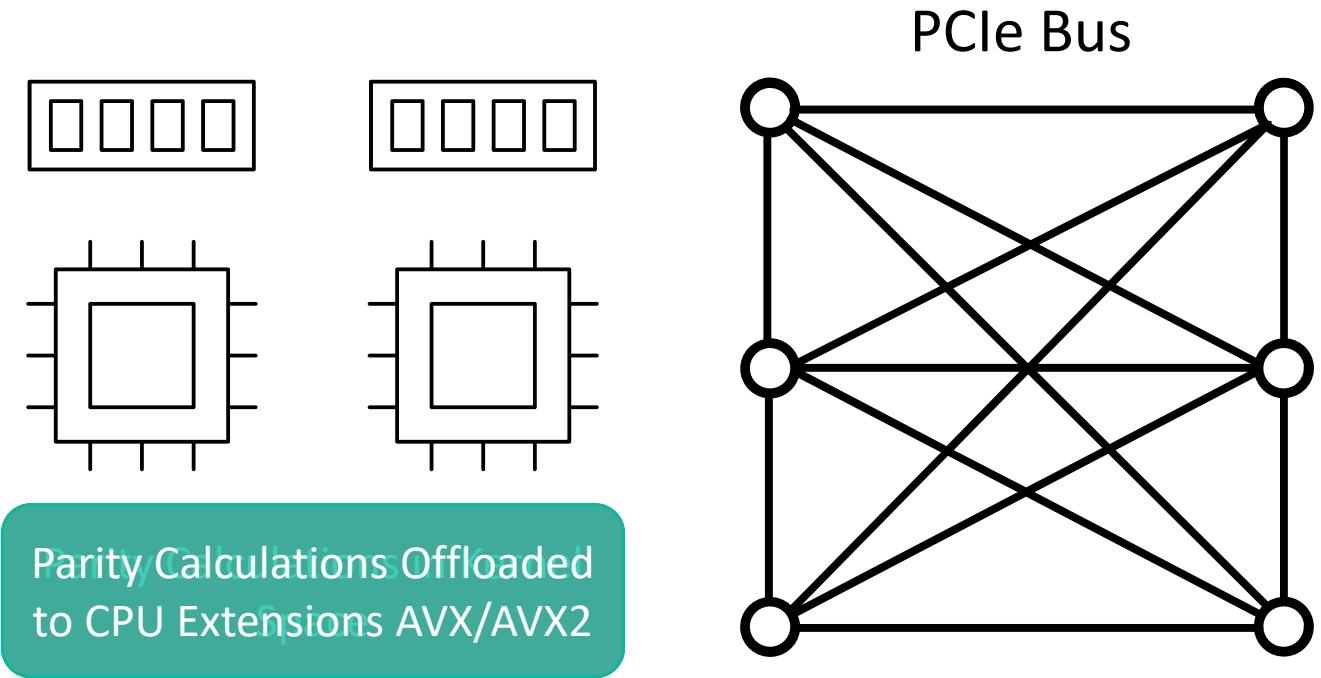


Networking

- 2 - 6 Ethernet Ports
- RoCE v2 or TCP
- RJ45 Management Port
- REST Based Management

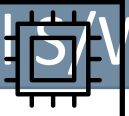


Hardware Accelerated Software RAID



FPGA Offload S/W RAID

GPU

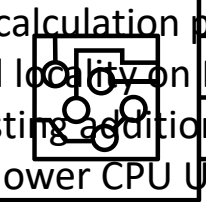


Traditional S/W RAID

- Parity calculation performed using FPGA
- Parity Calculations Offloaded to GPU
- Parity calculation performed in Kernel Space
- No H/W Offload
- High CPU Utilization, especially with small

GPU Offload S/W RAID

FPGA



- Parity calculation performed using GPU
- Shared locality on PCIe bus
- Interesting additional offload options
- Parity Calculations Offloaded to FPGA
- Much lower CPU Utilization & higher sustained write performance.
- Additional H/W cost



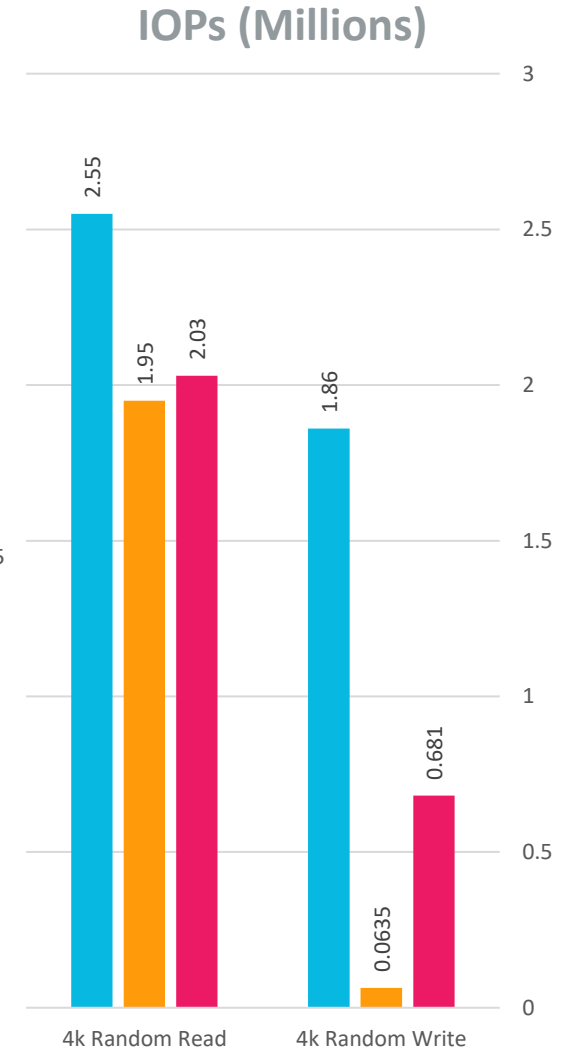
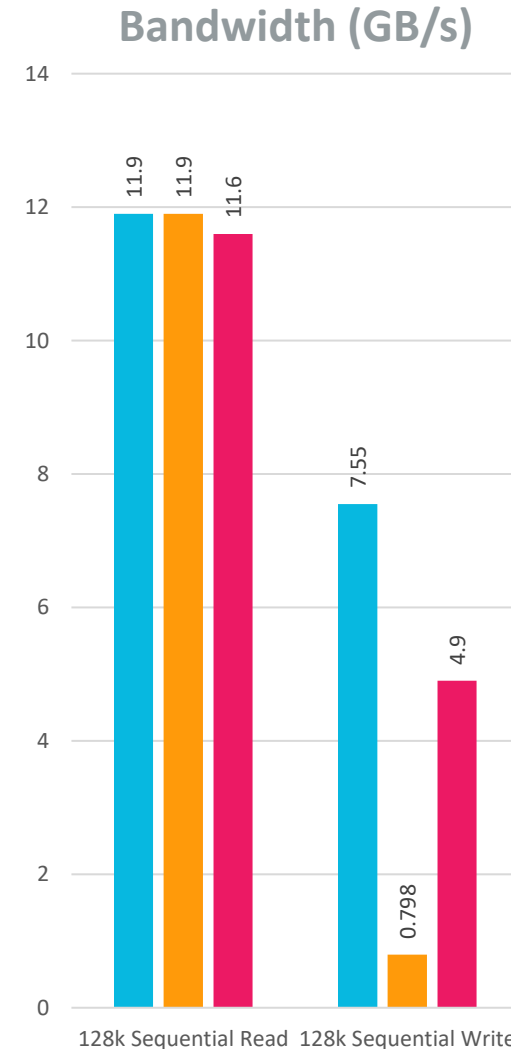
Hardware Accelerated Software RAID

Performance Example with RAIDIX ERA

- Test configuration:
 - 8x Dual Port NVMe Drives with single path.
 - 2x Lanes of PCIe Gen3 per drive
 - MDADM – 7+1 RAID 5
 - RAIDIX – 7+1 RAID5
 - Sequential 128k
 - Random 4k

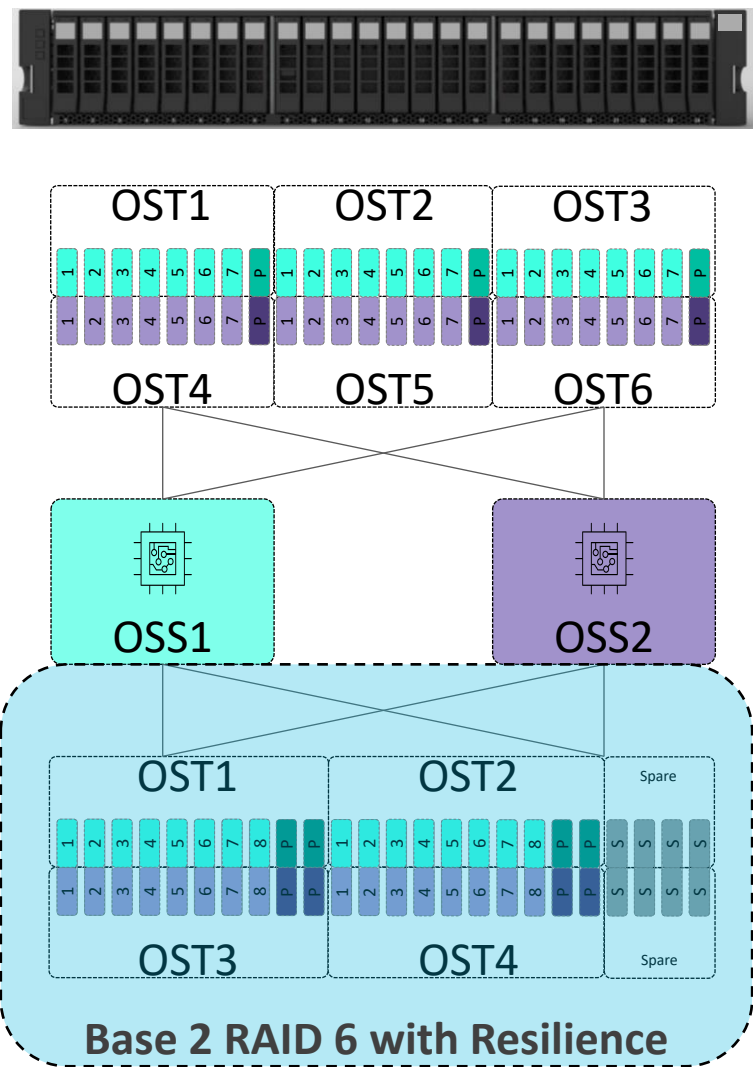


Note: Depending on stripe size, 128k block size may not be ideal for 7+1



HA Server Performance Planning

Existing Architecture



	Write Bandwidth (GB/s)		
	RAW	mdadm	RAIDIX
OST (7+1)	7.55	0.8 (11%)	4.9 (65%)
OST per OSS	3x	3x	3x
OSS	21.65	2.4	14.7

Summary

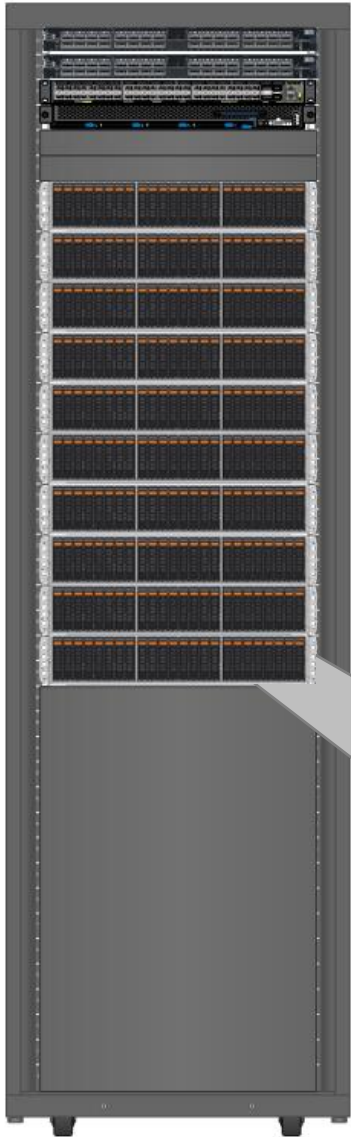
- 2U Dual Node server with 24 Dual Port NVMe™ SSDs
- 6 OST volumes built from 8 NVMe namespaces in a 7+1 RAID5
- OSS1 is Active on OST{1-3} and Passive on OST{4-6}
- OSS2 is Active on OST{4-6} and Passive on OST{1-3}
- Difficult to designed for full performance in failover condition
- Network performance needs 2 x 100 Gb links to support reads
- Poor write performance with mdadm at 11% of RAW

Challenges

- OST failover can only happen to other OSS node in same enclosure
- 24 SSD capacity limits potential number of base 2 RAID sets (2 x 8+2, 4 x 4+1)

Current Rack Diagram With HA NVMe Server

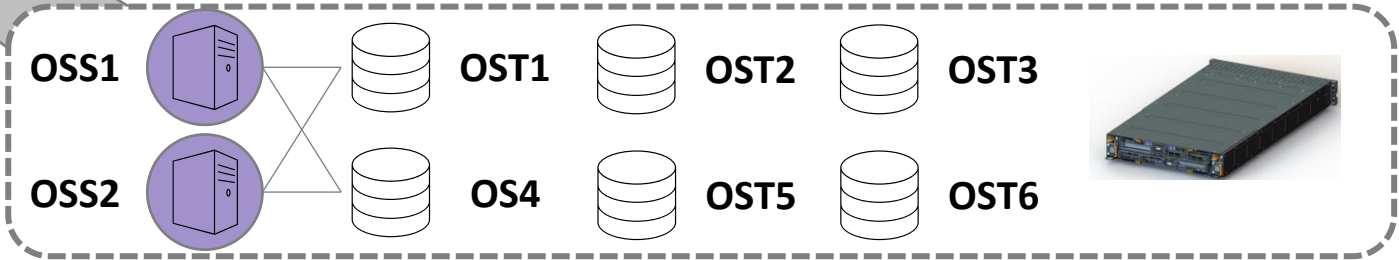
10 x Dual Node OSS Enclosures



	OST (mdadm)	Enclosure	RACK
HA Enclosures			10
#OSS		2	20
#OST		6	60
Power		1,600 W	17 kW
Write BW (GB/s)	0.8	2.4	24

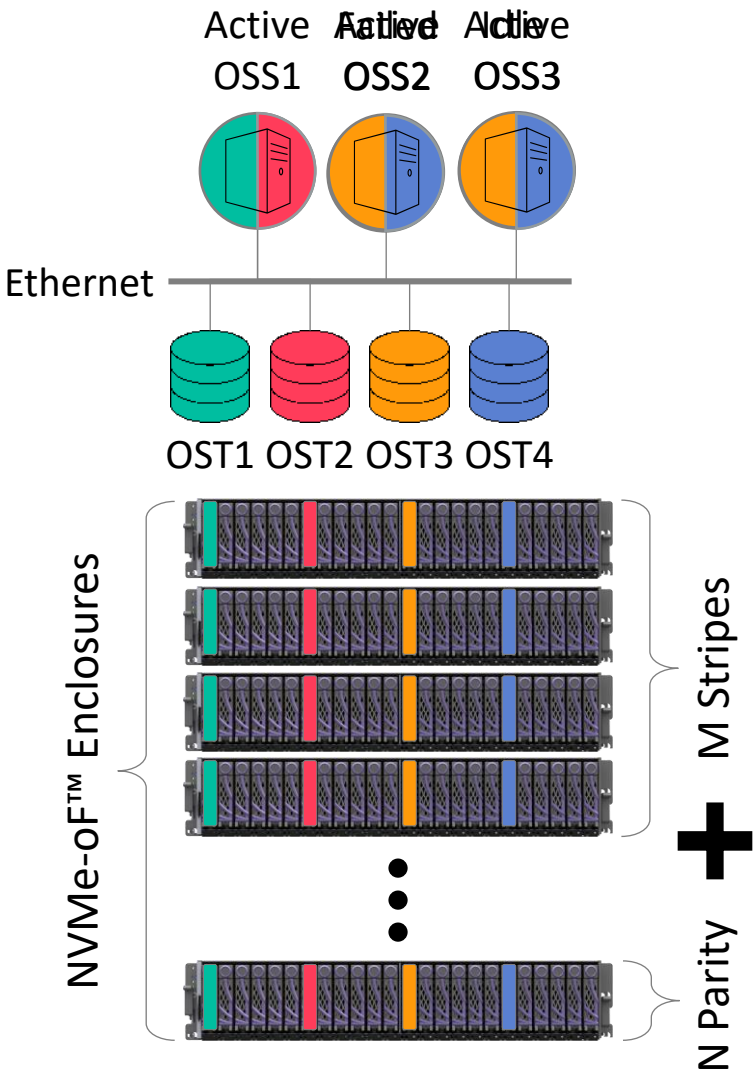
Primary Design Consideration

50% of compute & network resources are reserved for failover



NVMe-oF Architecture

Any-to-Any Topology



	Write Bandwidth (GB/s)			
	RAW	mdadm	Accel. RAID	Accel. B ₂ RAID
	8	7+1	7+1	8+1
OST	7.55	0.8 (11%)	4.9 (65%)	7 (93%)
OST per OSS	2x	2x	2x	2x
OSS	15.1	1.6	9.8	14

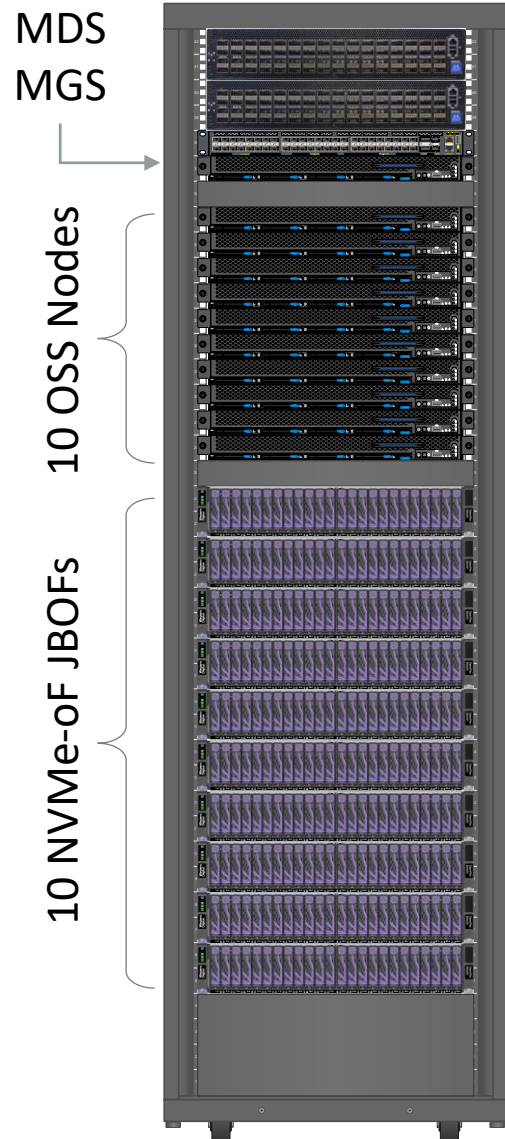
Principles

- Use 1 device from multiple enclosures to create a RAID stripe
- Stripe can be any RAID layout – 4+1, 7+1, 8+1, 8+2, m+n
- Any OSS can own any OST – No fixed pairing
- Idle OSS can take over for any other OSS failure
- Active OSS host as many OST volumes as capable – No 50% reserve
- Improve OST/OSS ratio through accelerated SW RAID

Benefits

- **Resilience:** Any SSD / Server / Enclosure / Network link can fail
- **Flexibility:** RAID not limited within a single HA node w/ 24 drives
- **Cost:** Reduced server over provisioning (i.e. reduced server cost)
- **Selection:** Server choice for OSS no longer needs to be HA node
- **Performance:** Improved write performance

NVMe-oF Rack Architecture



	HA Node Approach		NVMe-oF Approach	
	OSS	Rack	OSS	Rack
#JBODs				10
#OSS		20		8+2
#OST	3x 7+1	60	3x 8+2	24
Power		17kW		12kW
Write BW (GB/s)	2.4	24	21	168

20% of Compute resources are reserved for failover

50% Reduction in Server Count

7x Improvement in Sequential Write Performance

30% Reduction in Per Rack Power Requirement



Proof of Concept Results

9-Node Lustre Cluster with 8-Clients

Proof of Concept

Hardware

- 1x MDS
 - Platform: Dell® R650
 - Processor: 2x Intel® 5317 150TDP 12-Core 3.0GHz
 - Memory: 128GB (8x16GB 2933MHz)
 - Fabric: 1x ConnectX-6® 200 Gb Ethernet HCA
 - Storage: 10x 3.2TB WDC SN640 NVMe SSDs
- 8x OSS
 - Platform: Dell R650
 - Processor: 2x Intel 5317 150TDP 12-Core 3.0GHz
 - Memory: 128GB (8x16GB 2933MHz)
 - Fabric: 2x ConnectX-6 200 Gb Ethernet HCA
 - Storage: Remote NVMeoF
- 8x Clients
 - Platform: Dell R750
 - Processor: 2x Intel 6354 205TDP 18-Core 3.0GHz
 - Memory: 512GB (16x32GB 3200MHz)
 - Fabric: 1x ConnectX-6 200 Gb Ethernet HCA
- Networking:
 - SN3800 – 64-Port 100 Gb Switch
 - Storage Subnets 1 & 2
 - Lustre Subnet 1
 - SN2700 – 32-Port 100 Gb Switch
 - Lustre Subnet 2
- NVMe-oF Storage:
 - 3x Western Digital OpenFlex™ Data24 NVMe-oF™
 - 24x Ultrastar® DC SN840 3.2TB per Data24



Proof of Concept

Software

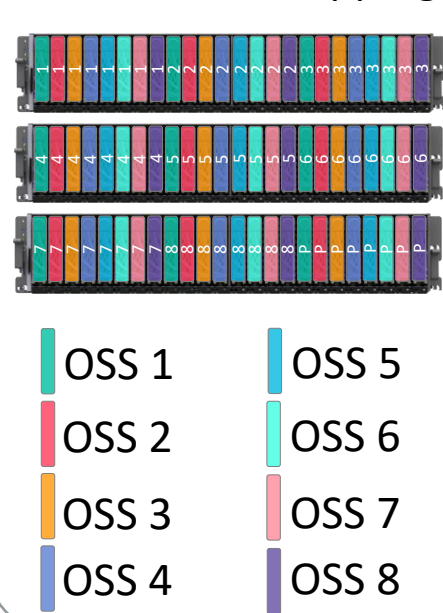
- MDS/OSS
 - Operating System: RHEL 8.3
 - Kernel: kernel-4.18.0-240.1.1.el8_lustre. x86_64
 - Network Stack: In-Box Mellanox 5.0.0
 - Lustre: Feature Release 2.14.0-1
 - RAID Software: RAIDIX ERA 3.3.0-289
- Clients
 - Operating System: RHEL 8.3
 - Kernel: kernel-4.18.0-240.22.1.el8_3.x86_64
 - Network Stack: In-Box Mellanox 5.0.0
 - Lustre: Feature Release 2.14.0-1
- Networking:
 - Storage:
 - RoCEv2 with Priority Flow Control
 - 2x Storage Subnets
 - Native NVMe Multipathing
 - Lustre:
 - RoCEv2 with o2ib
 - 2x Lustre Subnets



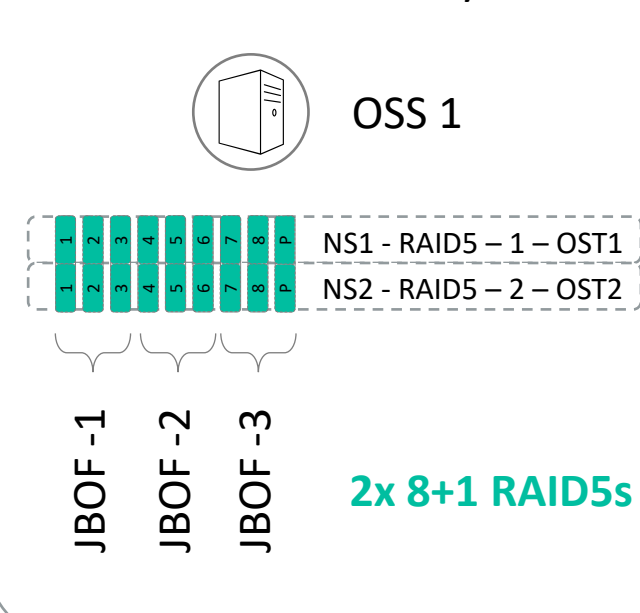
Proof of Concept

Architectural Diagrams

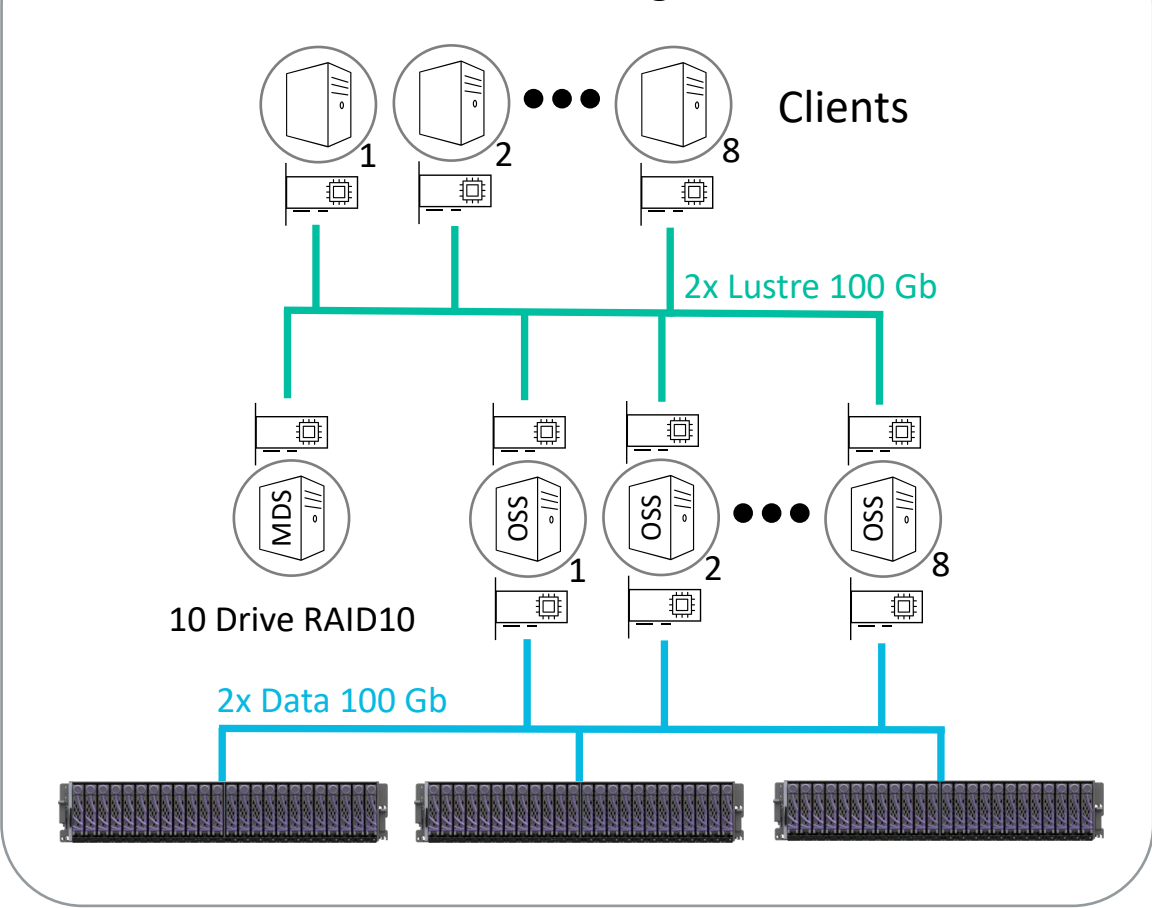
Drive to OSS Mapping



OSS RAID/OST Layout



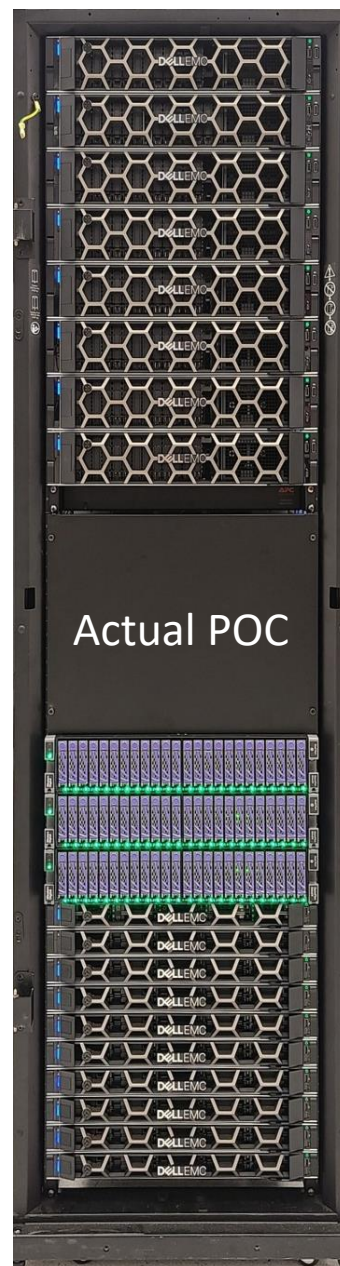
Network Diagram



Proof of Concept

Testing Methodology

- Benchmark 1: Flexible I/O Tester (fio)
 - File structure:
 - Lustre stripe count '-1'
 - Each client had its own directory
 - Each client had 48x 256 GB test files
 - Each file was in its own subdirectory
 - fio configuration:
 - IO Engine: libaio
 - 10 jobs per file
 - 128k sequential reads and writes
 - Queue depth of 16
 - DirectIO enabled
 - Testing Methodology:
 - Test Configurations:
 - Raw – Each OSS tests 18 Namespaces
 - RAIDIX – Each OSS tests 2 RAID Groups 8+1
 - Lustre – Fio tests as described above
 - Run tests 3 times and average tests
 - 2x 128k sequential fills
 - 1x 128k sequential writes (20 minutes)
 - 1x 128k sequential reads (20 minutes)



- Benchmark 2: Interleaved or Random (IOR)
 - File Structure:
 - Lustre stripe count '-1'
 - Each client had its own directory
 - Each client had 36x 512 GB test files
 - Each file was in its own subdirectory
 - IOR configuration:
 - MPI: OpenMPI
 - IO Engine: AIO
 - 1m sequential reads and writes
 - 288 Processes
 - DirectIO enabled
 - Collective IO
 - Reordered Tasks
 - 'fsync' on write close
 - Testing Methodology
 - 4 Iterations

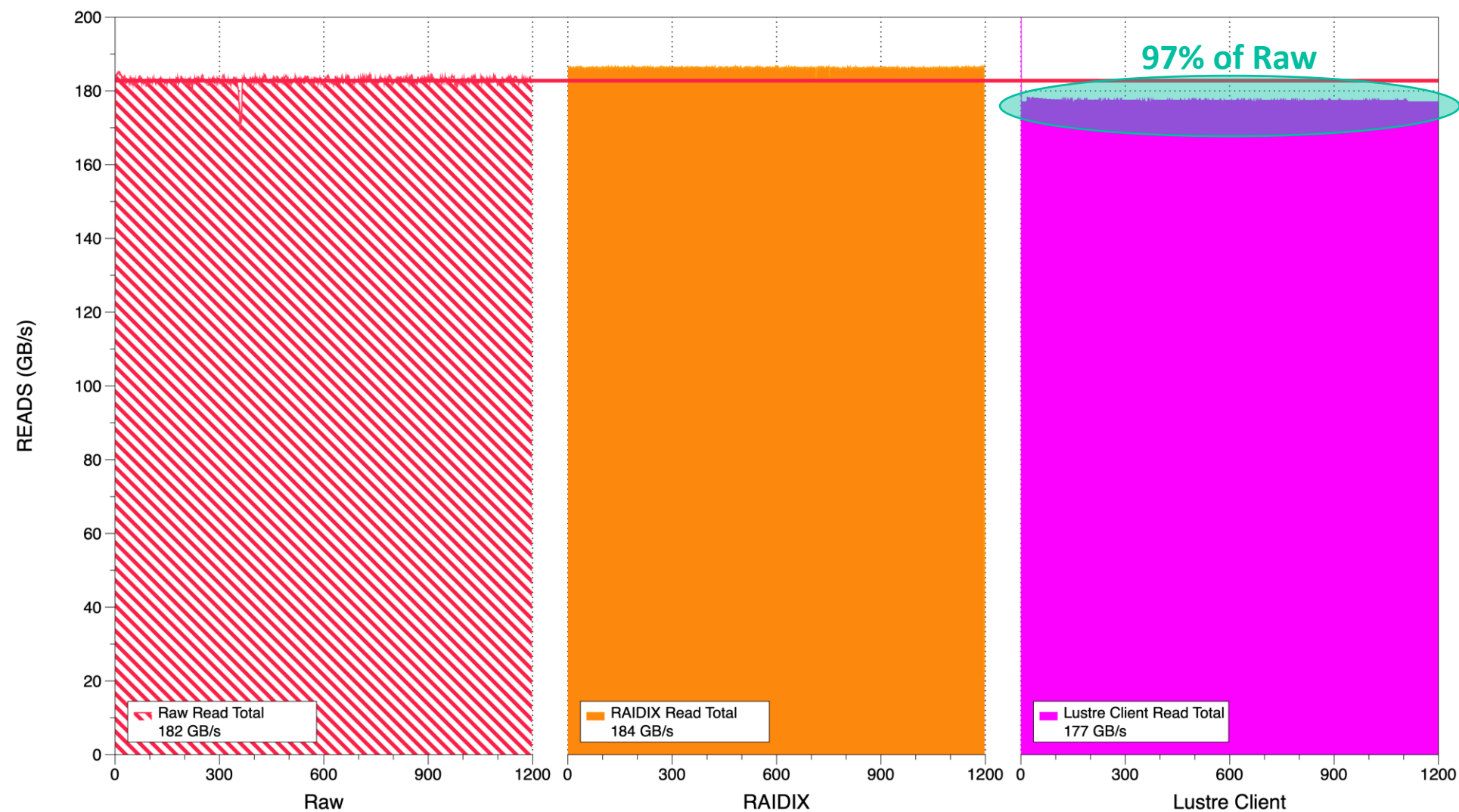
Proof of Concept

fio Results Summary

Test	BS	Raw	RAIDIX	Lustre FS
Sequential Write	128K	119.84 GB/s	112.14 GB/s	96.51 GB/s
Sequential Read	128K	182.13 GB/s	184.16 GB/s	177.36 GB/s
Sequential Write % from Raw	128K	100%	94%	81%
Sequential Read % from Raw	128K	100%	101%	97%

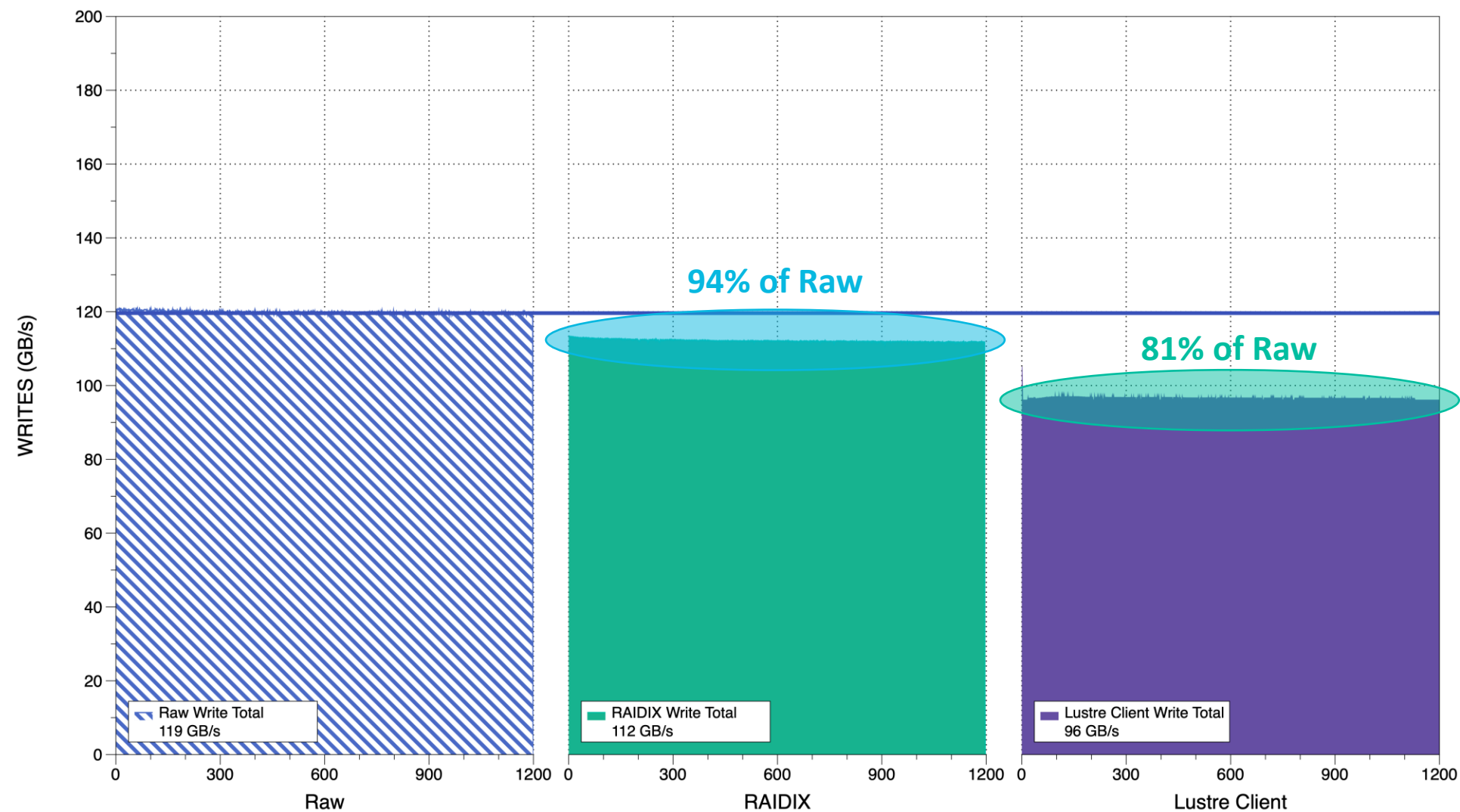
Proof of Concept

Time Series Read



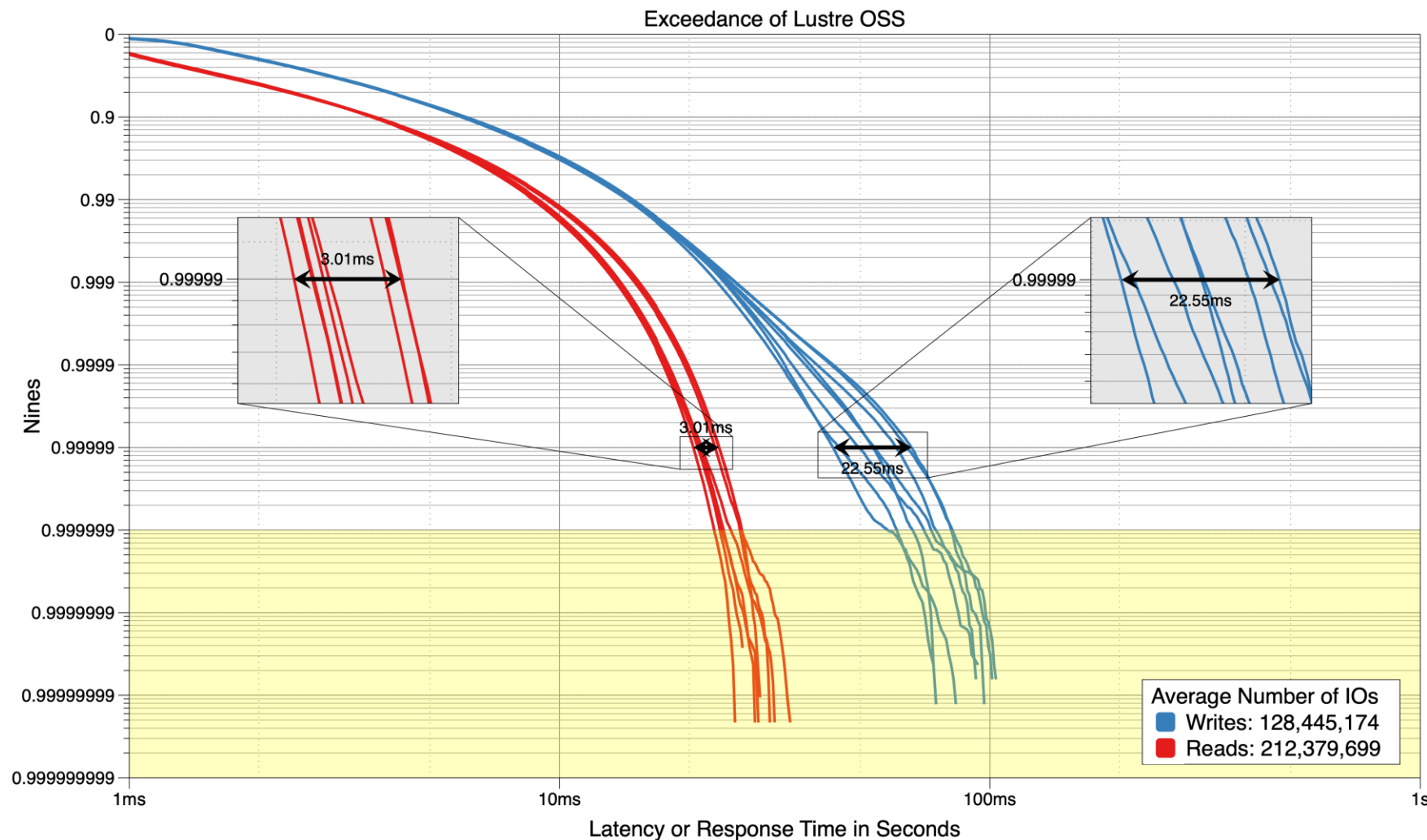
Proof of Concept

Time Series Write



Proof of Concept

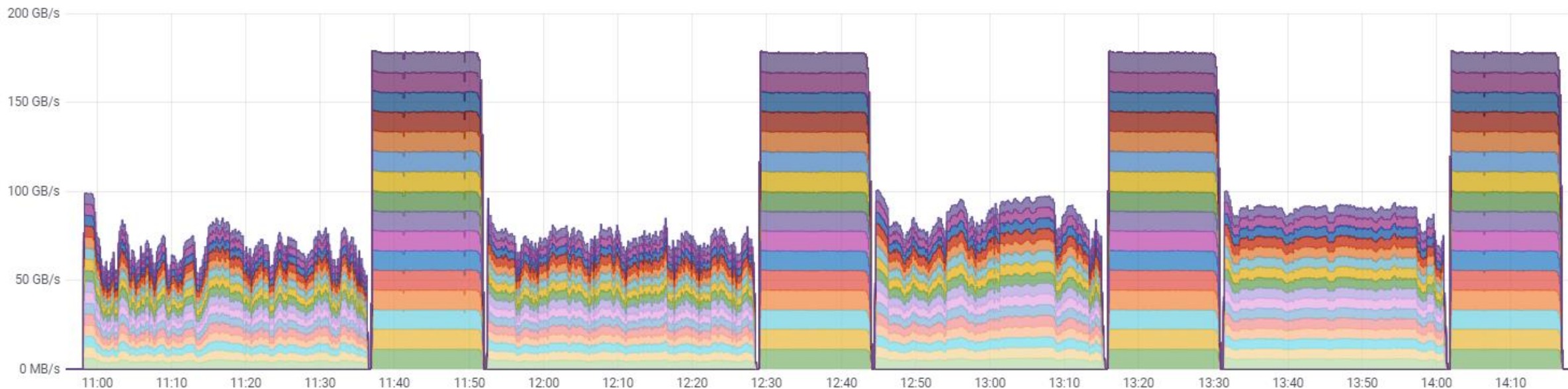
OSS Performance Variation



Nines: The percent of IOs completing in less than a given Latency or Response Time

Proof of Concept

IOR Results



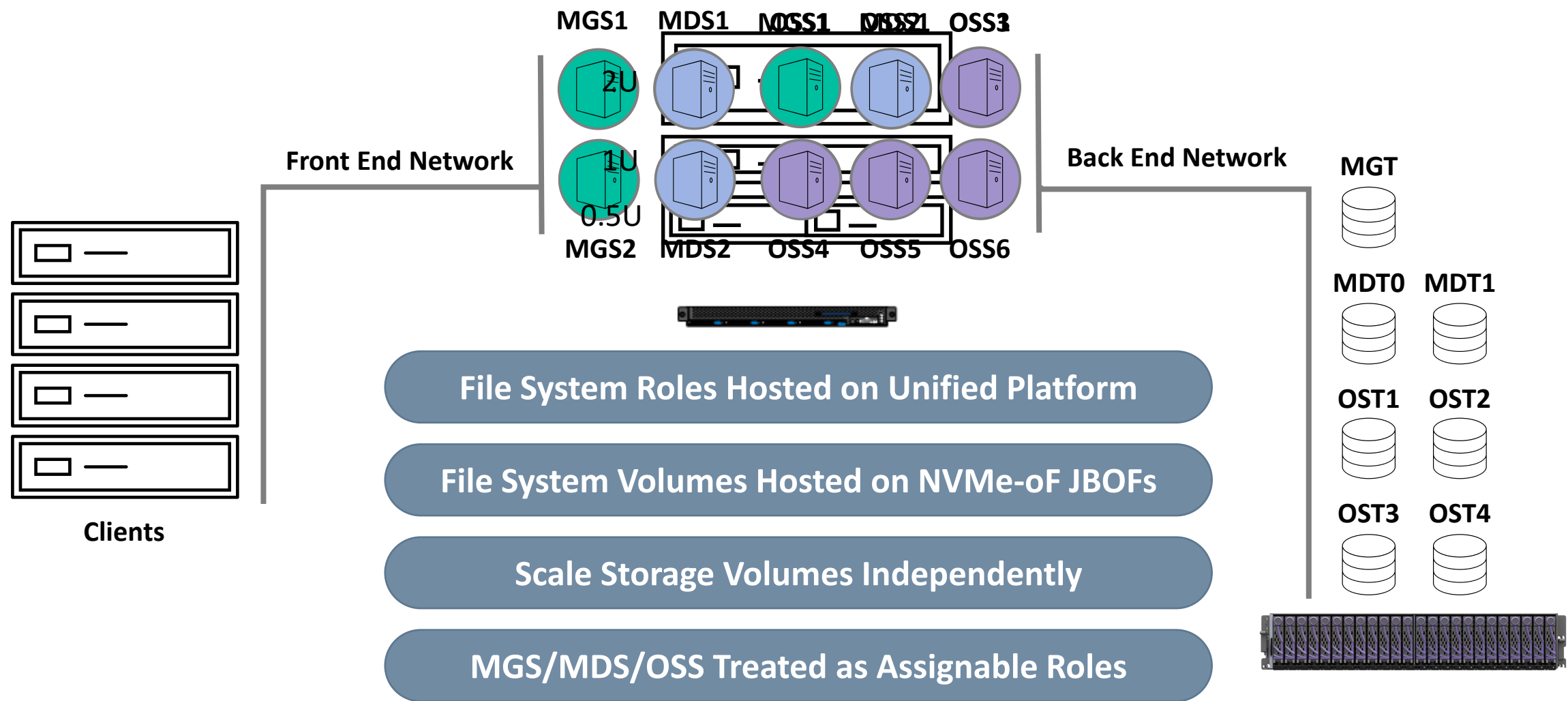
Summary of all tests:

Operation	Max(MiB)	Min(MiB)	Mean(MiB)	StdDev	Max(OPs)	Min(OPs)	Mean(OPs)	StdDev	Mean(s)
write	84001.51	65770.02	75301.53	7702.85	84001.51	65770.02	75301.53	7702.85	2026.59318
read	167585.55	167428.57	167512.84	58.27	167585.55	167428.57	167512.84	58.27	901.39336

79 GB/s

175 GB/s

NVMe-oF Parallel File System Architecture





Western Digital®

Western Digital, and the Western Digital logo are registered trademarks or trademarks of Western Digital Corporation or its affiliates. The NVMe and NVMe-oF marks are trademarks of NVM Express, Inc. PCIe is a registered trademark of PCI-SIG. All other marks are the property of their respective owners. Dell, the Dell logo, and other trademarks are trademarks of Dell Inc. or its subsidiaries Intel is a trademark of Intel Corporation or its subsidiaries in the U.S. and/or other countries. Mellanox and ConnectX are registered trademarks of Mellanox Technologies, Ltd.